A linked open data approach for geolinguistics applications

Emanuele Di Buccio, Giorgio Maria Di Nunzio and Gianmaria Silvello*

Department of Information Engineering,
University of Padua,
Via Gradenigo, 6/A Padova, 35131, Italy
Email: dibuccio@dei.unipd.it
Email: dinunzio@dei.unipd.it
Email: silvello@dei.unipd.it
*Corresponding author

Abstract: Geolinguistic systems explore the relationship between language and cultural adaptation and change and they can be used as instructional tools, presenting complex data and relationships in a way accessible to all educational levels. However, the heterogeneity of geolinguistic projects has been recognised as a key problem limiting the reusability of linguistic tools and data collections. We propose an approach based on LOD, which moves the focus from the systems handling the data to the data themselves with the main goal of increasing the level of interoperability of geolinguistic applications and the reuse of the data. We defined an extensible ontology for geolinguistic resources based on the common ground defined by current European linguistic projects. We provide a Geolinguistic Linked Open Dataset based on the data case study of a linguistic project named ASIt. Finally, we show a geolinguistic application, which exploits this dataset for dynamically generating linguistic maps.

Keywords: linked open data; digital geolinguistics; RDF; ontology; metadata.

Reference to this paper should be made as follows: Di Buccio, E., Di Nunzio, G.M. and Silvello, G. (2014) 'A linked open data approach for geolinguistics applications', *Int. J. Metadata, Semantics and Ontologies*, Vol. 9, No. 1, pp.29–41.

Biographical notes: Emanuele Di Buccio is a Post-doc Researcher at the Department of Information Engineering of the University of Padua, Italy, since 2011. He holds a PhD in Information Engineering from the Doctorate School in Information Engineering at University of Padua. His research interests include information retrieval and digital libraries. Since 2011, he has been working in the ASIt Project on the design and the development of a digital library system to support the analysis of geolinguistic variations in Italian dialects.

Giorgio Maria Di Nunzio is an Assistant Professor at the Department of Information Engineering in the University of Padua, Italy, since 2007. He holds a PhD in Computer Science from the University of Padua. His main research interests are machine learning, cross-lingual information retrieval and geolinguistic digital libraries. He has been involved since 2004 in numerous national and international projects about digital libraries and cross-lingual information access systems. Since 2010, he has been coordinating the computer science research unit of the multidisciplinary national project named ASIt.

Gianmaria Silvello is a Post-doc Researcher at the Department of Information Engineering in the University of Padua, Italy, since 2011 and he holds a PhD in Information Engineering from the Doctorate School in Information Engineering at University of Padua with a thesis on a set-based approach to deal with hierarchical structures. His main research interests are digital archives, data models, and experimental evaluation. Since 2006 he has been working on the design and development of a digital archive system in cooperation with the Italian Veneto Region. Since 2010 he has been working within the PROMISE European network of excellence.

1 Introduction

The research field of linguistics studies all aspects of human language, including morphology (the formation and composition of words), syntax (the formation and composition of phrases and sentences from these words) and phonology (sound systems) (Akmajian et al., 2010). Research in the variations in languages allows linguists to understand the fundamental principles that underlie language differences, language innovation and language variation in time and space. The

variation of a language is called dialect; and the study of these variations has led to the constitution of two research fields: dialectology and dialectometry. Dialectology is ultimately concerned with grammatical, lexical and phonological features that correspond to regional areas. Dialectometry mainly concentrates on the regional distribution of dialect similarities, which can be labelled according to a more or less slight variance of dialect between bordering locations.

Traditional studies in dialectology were generally aimed at producing dialect maps, whereby imaginary lines were drawn over a map to indicate different dialect areas. Since the end of the XIX century,[1] linguists have produced extensive cartographic work, most notably in the form of linguistic atlases (Lameli et al., 2010). Manuel Alvar López, one of the most important dialectologist of Spain, presented (Alvar Lòpez and Nuño Alvarez, 1981) an automated linguistic atlas which highlighted the advantages of a computerised versus manually drawn and reproduced atlas; the database developed for this atlas facilitated mapping on-demand and preparation of indices used in the interpretation of linguistic atlases.

In the last 30 years, modern GIS have provided efficient analysis of spatial data in many fields. "Geolinguistics is an interdisciplinary field that incorporates language maps depicting spatial patterns of language location or the results of processes that lead to language change" (Hoch and Hayes, 2010). The synergy between geography and linguistics is well-outlined by Breton, who described the process through which a geographic thought becomes a tool for linguists (Breton and Schiffman, 1991):.

> In analysing the distribution in space and in society of the facts of language, the linguist employs the methods of geography, cartography and the establishment of correlations and causalities between spatial phenomena.

In this context, the linguistic atlas has proved to be a vital tool and product of geolinguistics since the earliest stages of the field, and it has provided a stage for the incorporation of modern GIS.

In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide. The World Atlas of Languages Structures (WALS) (Haspelmath et al., 2005) is one of the largest projects, with 160 maps showing the geographical distribution of structural linguistic features,[2] and it is the first linguistic feature atlas on a worldwide scale. The study of dialectal heritage is the goal of many research groups in Europe as well. The aim of the EU-sponsored Common Language Resources and Technology Infrastructure project (CLARIN; Odijk, 2010) is to create an infrastructure which makes language resources (annotated recordings, texts, lexica, ontologies) and technology (speech recognisers, lemmatisers, parsers, summarisers, information extractors) available and readily usable to scholars of all disciplines. The Edisyn Project (Kunst and Wesseling, 2011) is a project on dialect syntax funded by the European Science Foundation, the goal of which is to establish a European network of researchers using similar standards with respect to methodology of data collection, data storage and annotation, data retrieval and cartography.

The heterogeneity of linguistic projects has been recognised as a key problem limiting the re-usability of linguistic tools and data collections (Chiarcos, 2012). The rate of re-use for linguistic database technology together with related processing tools and environments is still too low. A relevant example of how heterogeneity negatively affects the re-use and integration of data is represented by the Edisyn search engine – the aim of which was to make different dialectal databases comparable – which 'in practice has proven to be unfeaseable'.[3] In order to find common ground where linguistic material can be shared and re-used, the methodological and technological boundaries existing in each research linguistic project need to be overcome.

The research direction we pursue in this work is to move the focus from the systems handling the linguistic data to the data themselves. To this end, the LOD paradigm (Heath and Bizer, 2011) represents the most natural choice, because it eases interoperability between different systems by allowing the definition of data-driven models and applications. LOD is based on the definition of real-world objects, identified by means of a dereferenceable URI.[4] Objects are related to one another by means of typed links. Interoperability is achieved by a unifying data model (i.e. RDF[5]), a standardised data access mechanism (i.e. HTTP), hyperlink-based data discovery (i.e. URI), and self-descriptive data (based on shared open vocabularies from different namespace) (Heath and Bizer, 2011).

Building on LOD methodologies and technologies, we present

- an extensible OWL-based model for describing geolinguistics resources, the purpose of which is to enable interoperability at a data-level by overcoming the single collection characteristics;
- the ASIt linguistic project which is based on micro-variations of Italo-Romance dialects; we report on the design of the conceptual and logical model of the database;
- the mapping between the ASIt model and the newly defined geolinguistics OWL-based model along with the curated linguistic linked open data set generated from the original ASIt database;
- a geolinguistic web application that provides functionalities for accessing, browsing, searching the data by means of linguistic features, and visualising the data on dynamically generated maps.

The paper is organised as follows: Section 2 presents the LOD principles and technologies and describes current linguistic projects. In Section 3, we define the ontology for representing geolinguistic resources and we describe its realisation using OWL. Section 4 describes the ASIt enterprise and reports on the rules for mapping the ASIt conceptual model into the newly defined ontology; Section 5 presents the architecture of the ASIt system composed of the linguistic layer and the RDF layer. The publicly available linked open data set is presented in Section 6 and the web application to access it is described in Section 7. Finally, Section 8 draws some conclusions and discusses future work.

2 Background

2.1 Linguistic resources

Language resources that have been made publicly available can vary in the richness of the information they contain: on the one hand, a corpus typically contains at least a sequence of words, sounds or tags; on the other hand, a corpus may contain a large amount of information about the syntactic structure, morphology, prosody and semantic content of every sentence, plus annotations of discourse relations or dialogue acts (Bird et al., 2009). However, the quality of such corpora may have been reduced by the intense, and often poorly controlled, usage of automatic learning algorithms (Spärck Jones, 2007). Depending on the type of analysis a researcher performs, linguistic data sets created by an automatic PoS tagger can be either helpful or useless. For example, PoS annotations are very important for performing shallow linguistic analyses on large corpora, while they are not suitable for capturing fine-grained grammatical differences. Indeed, when comparing various dialectal translations of the same sentence, even an accuracy of 98% of the best automatic PoS tagger is not sufficient to pin down these subtle asymmetries. This specificity can only be achieved manually (Agosti et al., 2010).

Tagging is one of the necessary steps required to prepare a linguistic resource (Kilgarriff, 2007), and it requires human intervention to achieve the highest quality necessary for re-usable linguistic data. Curated databases[6] (Buneman et al., 2008) are a possible solution for designing, controlling and maintaining collections that are consistent, integral and high quality. To this end, Bird et al. (2009) discuss three important points about the design and distribution of language resources:

- How do we design a new language resource and ensure that its coverage, balance and documentation support a wide range of uses?
- When existing data is in the wrong format for some analysis tool, how can we convert it into a suitable format?
- What is a good way to document the existence of a resource we have created so that others can easily find it?

We address these issues by adopting an approach based on the LOD paradigm with the aim of enabling interoperability at a data level.

2.2 Linguistic projects

Current linguistic projects are usually general collections of language resources. They often provide interoperability at a system level by harvesting linguistic data from different sources. The OLAC,[7] which recently celebrated its first 10 years of activity, is a worldwide network dedicated to collecting information on language resources (field notes, grammars, audio/video recording, descriptive papers, and so on) and developing standard protocols for interoperability. GOLD[8] was the first ontology to be designed specifically for linguistic description on the semantic web (Farrar and Langendoen, 2003). It proposes a solution to the lack of interoperability between linguistic projects and projects designed specifically for NLP applications. It can act as a kind of lingua franca for the linguistic data community, provided that data providers are willing to map their data to GOLD or to some similar resource. Chiarcos et al. (2008) present a framework for producing multi-layer annotated corpora: a pivot format serving as 'interlingua' between annotation tools, an ontology-based approach for mapping between tag sets, and an information system that integrates the various annotations and allows for querying the data either by posing simple queries or by using the ontology.

The research on morphology and syntax of regional varieties of languages requires much more fine-grained data than NLP applications and the relevant material requires manual elicitation. In the context of the Edisyn European network, there are research projects, such as the Dynamic Syntactic Atlas of the Dutch Dialects (DynaSAND), the 'Syntax-oriented Corpus of Portuguese Dialects' (CORDIAL-SIN), the Nordic Dialect Corpus (ScanDiaSyn), which have the aim of forming a solid basis for dialectological and other sociolinguistic studies. DynaSAND is an online tool for linguistic research (Bael et al., 2006, Chapter 4, pp.54–90). It consists of a database, a search engine, a cartographic component, and a bibliography on dialect syntax. The CORDIAL-SIN[9] project aims at making available a significant amount of spontaneous and semi-directed speech drawn from the collected data. This project also aims at providing fast and systematic access to precise morphological and syntactic information. ScanDiaSyn[10] is a corpus of Norwegian, Swedish, Danish, Faroese, Icelandic and Övdalian spoken languages. It consists of spontaneous speech data from dialects of the North Germanic languages.

2.3 Geolinguistic projects

In November 2012, the 'Symposium of 20 years of Geolinguistics'[11] was organised to celebrate the anniversary of one of the most important digital geolinguistic projects: VIVALDI (Vivaio Acustico dell Lingue e Dialetti d'Italia).[12] The symposium gathered together the most important European digital geolinguistic projects. Since these projects have been running for years, in this section we present their objectives along with the URL of their online web page. We do not present ASIt here, even though it was present at the symposium.

The *ALD: linguistic atlas of dolomitic ladinian and neighbouring dialects* project[13] belongs to the classic tradition of linguistic geography in romance studies. It considers the speech community of northeastern Italy and southeastern Switzerland. It collects and documents the dialectal competencies of essentially multilingual informants who live in one of the 217 surveying points of this community. The data, which has been collected, makes it possible to examine the dialectal area.

AMPER: multimedia atlas of the prosody of the Romanic space[14] is a project with the aim of describing and

characterise the prosody of the variety of Romanic languages around the world, in particular the Iberian Pensinsula and South America. The main focus is a phonetic study of the language, with phonological, dialectological and sociolinguistic implications. The project allows the study of the results visualised on maps that can be searched and examined online.

The project *Atlas of everyday German language*[15] aims at studying colloquial German and the current diversity of the German language. The data, gathered by means of online questionnaires and surveys, are displayed on maps and allow the comparison between old German words and the corresponding new German expressions.

The objective of the *ALAVAL: THE Atlas Linguistique Audiovisuel du Francoprovençal Valaisan* project[16] is to record the particular Francoprovençal dialects spoken in a small area of western Switzerland. The peculiarity of this project is that it stores data that combines language and gesture, verbal and non-verbal communication. In particular, the multimedia data consist of recordings of the speakers which capture phenomena of spontaneous orality (hesitation, reformulation, etc.) that are not present in other projects.

The *soundcomparison*[17] project aims at recording the variety of the sounds of the English language and two varieties of the two main indigenous language families spoken in the Andes, on various levels: *in time*, with transcriptions of historical ancestor forms of English; and by *sociolinguistic context*. The project allows the search on interactive maps of 'instant playback' recordings of variation in phonetics across the regional dialectal/accent diversity within any given language family.

The *dialectal corpus of extremadura*[18] is constituted by a collection of oral and written documents. The fundamental objective of this project is to build a vast corpus of speech samples to analyse the current situation of linguistic modalities of the region of Spain named Extremadura. The online database, allows researchers to search more than 400 linguistic and ethnographic maps.

In this context an important project, which was not present at the symposium, is *LL-MAP*.[19] This project is designed to integrate language information with data from the physical and social sciences by means of a GIS (Xie et al., 2009). The most important part of the project is a language subsystem, which relates geographical information on the area in which a language is or has been spoken to data on resources relevant to the language. The system also includes information on topography, political boundaries, demographics, climate, vegetation and wildlife, thus providing a basis upon which to build hypotheses about language movement across territory.

The LL-MAP system encourages collaboration between linguists, historians, archaeologists, ethnographers and geneticists, as they explore the relationship between language and cultural adaptation and change. As a GIS, LL-MAP has the potential to be a captivating instructional tool, presenting complex data in a way accessible to all educational levels.

2.4 Linked open data

The LOD paradigm refers to a set of best practices for publishing data on the web[20] and it is based on a standardised data model, the RDF. RDF is designed to represent information in a minimally constraining way and it is based on the following building blocks: graph data model, URI-based vocabulary, data types, literals, and several serialisation syntaxes. The basic structural construct of RDF is a triple (subject, property, and object) which can be represented in a graph; the nodes of this graph are subjects and objects and the arcs are properties. Nodes and arcs are identified by URI. RDF adopts a property-centric approach allowing anyone to extend the description of existing resources; properties represent relationships between resources, but they may also be thought of as attributes of resources, like traditional attribute-value pairs.

In order to provide a mechanism to describe resources and properties, the RDF/S has been introduced. RDF/S provides "*basic elements for the description of ontologies intended to structure RDF resources*" (Klyne and Carroll, 2004). Basically, RDF/S defines classes, sub-classes, properties, and sub-properties in a hierarchical way. RDF distinguishes between classes and members of a class and we can specify the domain and range of each property. RDF/S defines the *intensional* level of the model, whereas the instances of the schema (the RDF triples) represent the *extensional* level. The RDF/S basic semantics do not always suffice to represent the reality of interest, for this reason OWL[21] has been introduced.

In this paper, we describe the ontology for representing geolinguistic resources and its realisation using OWL.

3 An extensible ontology for geolinguistic resources

The common ground defined by current European linguistic projects allows us to infer the fundamental classes and properties necessary to define an ontology for modelling and representing geolinguistic resources. Geolinguistic concepts can be organised into three major areas: geographic, derivation, and tagging. The geographical area comprehends classes and properties related to physical places. The derivation area is about people speaking a certain language, their relationships, and the geographical area where they live. Furthermore, the derivation area allows for the study of the correlation between social factors, education and knowledge of the dialect, and the distinctiveness of a regional dialect. Finally, the tagging area regards language-specific classes and properties, such as documents, sentences, words, and their relationships.

These three areas allow for a comprehensive description of geolinguistic resources; these main areas along with the defined classes and properties are shown in Figure 1 which is an updated and revised version of the diagram illustrated in the work of Di Buccio et al. (2012, 2013). The main classes of the geographical area are `region`, `province`, and `town` where a region is composed of one or more provinces, and a province

is composed of one or more towns. Since the geographical area defines the places where a dialect is spoken, it is connected to the derivation area by the properties relating the `town` and the `actor classes`. The `actor` class along with the connected `dialect class` allows us to model the people – hereafter named speakers – speaking a given dialect. Furthermore, the connection with the `town` class describes where the people live. The `hasAncestor` property of the `actor` class models the relationships of the people with their own ancestors; this property allows for a deeper analysis of interactions between speakers and the territory where a dialect is spoken. The actor class is central because it is related also to the main class of the tagging area, the `document` class. To this purpose, the `ActorSpeaks Document` class relates a document to the dialect and to the speaker.

A document represents the composite unit of study of a dialect; it is composed of one or more sentences which are subsequently divided in words for further analysis. A document may be redacted in one language (e.g. Italian or English) and then translated into several dialects which allow for linguistic comparisons. The syntactical analyses of these parallel translations are possible thanks to the `tag class` specialised into two main sub-classes: `sentence tag` and `PoS tag`. `Sentence tag` allows us to capture a sentence-level phenomenon, whereas `PoS tag` allows us to capture a phenomena occurring on a `word` in a `collocation`, i.e. a specific position within a given sentence. The `WordSentenceCollocation` class relates a tag to a word within a sentence along with the properties relating it to the `word`, `sentence` and `collocation` classes. The `SentenceDocument Collocation` class relates a sentence to a document specifying the collocation of the sentence within the document by means of the class `Sentence Collocation`.

As far as the vocabulary adopted in this specification is concerned, we use the namespaces and prefixes reported in Table 1; `asit` is the only vocabulary which is not inherited from other domains. In Table 2, we report the OWL data type properties of the described classes. We adopt OWL to specify the maximum number of occurrences of a class within a property.[22] The OWL specification of the geolinguistic ontology is publicly available at the following URL: http://purl.org/asit/rdf/asit-schema.rdf

This ontology is the starting point for modelling and describing geolinguistic resources because:

- it provides general-purpose concepts and relationships;
- it is extendable by adding more fine-grained classes;

it permits an easy mapping from existing linguistic projects and publicly available databases.

Figure 1 Diagram representing the RDF/S of the geolinguistic ontology (see online version for colours)

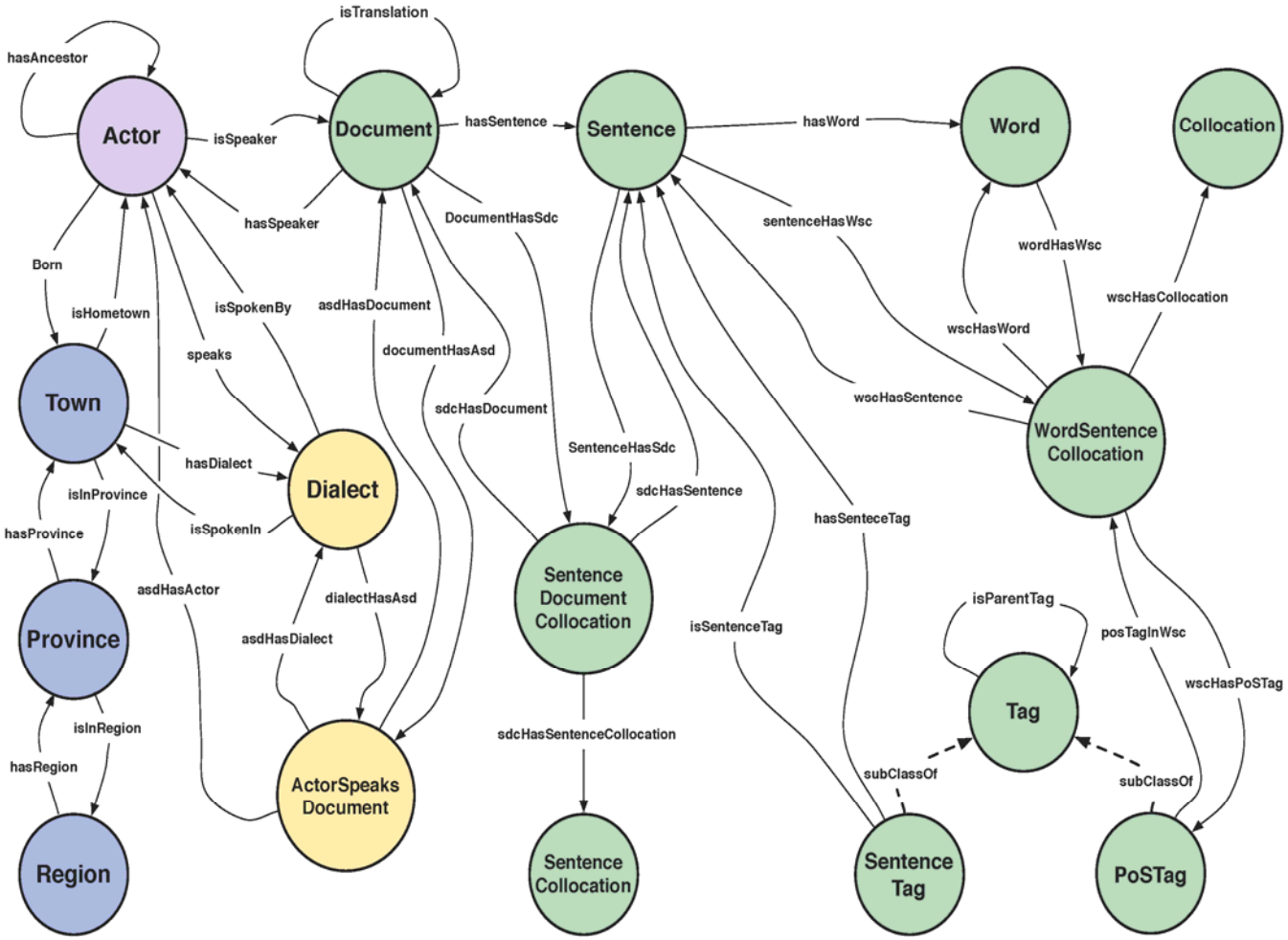


Table 1 Namespaces and prefixes adopted by the RDF specification

Prefix	*Namespace*	*Description*
asit	http://purl.org/asit/terms/	ASIt vocabulary terms
dcterms	http://purl.org/dc/terms/	Dublin Core terms
foaf	http://xmlns.com/foaf/0.1/	Friend of a friend
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	WGS84 Geo Positioning
gn	http://www.geonames.org/ontology#	GeoNames Ontology
owl	http://www.w3.org/2002/07/owl#	OWL vocabulary terms
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF vocabulary terms
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema

Table 2 Main data type properties of the classes of the schema of Figure 3

Area	*Class*	*OWL Datatype Properties*
Geographical	Region	gn:officialName, asit:geographicPartition, asit:regionNotes
	Province	gn:officialName, gn:shortName, asit:provinceNotes
	Town	gn:officialName, geo:alt, geo:lat, geo:long, gn:population, asit:townNotes, asit:provinceCapital, asit:provinceLittoral, asit:altimetricArea, asit:mountainTown, asit:surface, asit:latitudine, asit:longitudine
	Dialect	asit:dialectName
Derivation	Actor	foaf:firstName, foaf:lastName, foaf:name, foaf:birthday, foaf:gender, foaf:mailbox, asit:placeOfBirth, asit:education, asit:job, asit:country, asit:lang, asit:actorNotes, asit:affiliation
Tagging	document	dcterms:title, dcterms:date
	sentence	asit:sentence, asit:transcription, asit:sentenceNotes
	word	asit:wordText, asit:transcription
	collocation	asit:position
	sentenceCollocation	asit:sentencePosition
	tag	asit:tagDescription, asit:mandatory

4 Use case based on Italian dialects: ASIt

4.1 The ASIt enterprise

The ASIt enterprise builds on a long standing tradition of collecting and analysing linguistic corpora, which has originated different projects over the years (Agosti et al., 2010; Agosti et al., 201; Agosti et al., 2012). Dialectal data stored in ASIt were gathered during a 20-year-long survey investigating the distribution of several grammatical phenomena across the dialects of Italy (Benincà and Poletto, 2007). These data were collected by means of written questionnaires formed by sets of Italian sentences: speakers were asked to translate them into their dialects; therefore, each questionnaire is associated with many parallel dialectal translations. At present, there are eight different questionnaires written in Italian and almost 500 translations in more than 240 different dialects, for a total of more than 54,000 sentences, and more than 40,000 tags.

There are two aspects that characterise ASIt and make it different from other linguistic projects: the nature of Italian dialects and the kind of linguistic theory ASIt aims to interact with. The Italian dialectal area presents particular aspects of syntax, morphology and phonology. The kind of information ASIt wants to gather involves not only the presence of a certain linguistic feature, but also the absence of it; for example, the absence of the subject in a sentence. The absence of a term in conjunction with other specific linguistic features, differentiate the structure of one dialect from another.

There are many reasons why ASIt did not use state-of-the-art PoS automated programs. First, the 'trivial' identification of basic POS tags (nouns, verbs, etc.) is not enough to capture minimal cross-linguistics differences between closely related languages. Second, the linguistic variants cannot be reduced to lexical distinctions only, i.e. syntactic differences are in general unpredictable on the basis of the properties of single lexical items. Third, there are not enough training documents to tune the performance of a POS tagger for a specific dialect. Therefore, ASIt has adopted a fully manual tagging of the sentences and proposed the creation of a specific set of linguistic tags, starting from a universal core shared by all languages (on the basis of the work done by DynaSAND; Bael et al., 2006, Chapter 4, pp.54–90), and subsequently developing a language-specific periphery which is compatible with other projects.

4.2 Fine grained resource of linguistic data

The design and construction of the linguistic database has followed a three phase approach (Di Buccio et al., 2012): (a) the world of interest was represented at a high level by means of a conceptual representation based on the analysis of requirements; (b) it was progressively refined to obtain the logical model of the data of interest; (c) the digital library system and a web application to access the data were implemented.

A simplified view of the conceptual schema is reported in Figure 2. In this schema, we can highlight the three main conceptual areas of the geolinguistic ontology defined in Section 3:

- The *geographical area*, which is the place where a given dialect is spoken and where a speaker is born;
- the *derivation area*, which focuses on the background of the speaker: the level of knowledge of the dialect, the particular variety of the dialect, the birthplace, the ancestors, the document translated by the speaker;
- the *tagging area*, which is how the document is structured and how it has been tagged at sentence and word level.

In Table 3, we list the attributes of each entity for each area:

- *Geographic area*: the entity *region* has a GeographicPartition attribute which indicates a cardinal direction (north, south, north-east, etc.). It represents the geographical position of the region within a country. For example, it is used by linguists to gather information about 'southern' dialects. The entity *town* contains geographical information about the elevation above the sea, the surface of the town, the population and the geographical position. These attributes are used to make hypotheses about the influence of the morphology of a region, natural barriers such as mountains or rivers, on the evolution of dialects.
- *Derivation area*: the attributes of the entity *actor* are used to study the correlation between social factors, education and knowledge of the dialect. The place of birth and where the speaker currently lives are indicators, together with the level of knowledge of the dialect, of the distinctiveness of a dialect of a region.
- Tagging area: the entity *document* is a 'container' of sentences. In ASIt, it is a questionnaire prepared by linguists. The date of compilation of the questionnaire is used for diachronic studies of the dialects. The entity *tag* has an attribute type which distinguishes between sentence or part-of-speech.

4.3 The mapping between ASIt conceptual model and the ontology

We mapped the ASIt conceptual model into the geolinguistic ontology. To this purpose, we took into account several approaches (Auer et al., 2009; Myroshnichenko and Murphy, 2009) that have been proposed to map ER conceptual schemas into OWL-based models by employing a predefined set of rules. Let E a set of entities and R a set of relationships, we consider the following mapping rules:

- each entity $e_i \in E$ is mapped into a class (`rdfs:Class`) c_i;
- each binary relationship $r_{i,j} \in R$ (relating the $e_i \in E$ to $e_j \in E$) is mapped into a property (`rdf:Property`) $p_{i,j}$ with c_i as subject and c_j as object;
- for each ternary relationships $r_{i,j,k}$ (relating the entities $e_i, e_j, e_k \in E$) four `rdfs:Class` $c_i, c_j, c_k,$ and c_{ijk} are created. Then c_{ijk} is connected to c_i, c_j and c_k by three `rdf:Property` – i.e. $p_{i,ijk}$, $p_{j,ijk}$, and $p_{k,ijk}$;
- the attributes of an entity are considered data type properties and converted into RDF literals;[23] the data type of a given attribute is maintained by defining a proper XML Schema Datatype (XSD)[24] for the RDF literal.

Figure 2 A simplified entity-relationship schema of the data resource of Italian dialects: the three main areas of interest are highlighted with different background colours. Attributes of the concepts have been removed for better readability (see online version for colours)

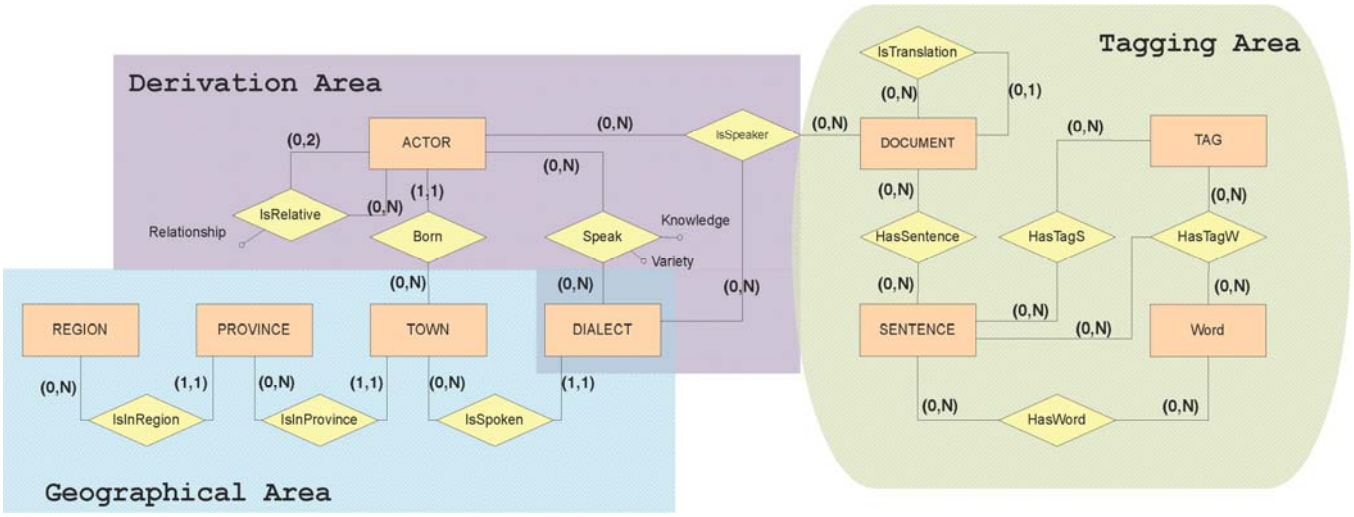


Table 3 Partial list of the attributes of the entities of the schema of Figure 2

Area	*Entity*	*Attributes*
Geographical	Region	Id, Name, GeographicPartition
	Province	Id, Name
	Town	Id, Name, Area, Elevation, Population, Latitude, Longitude
	Dialect	Id, Name
Derivation	Actor	Id, FullName, DateOfBirth, Sex, Education, Job
Tagging	Document	Id, Title, DateOfCompilation
	Sentence	Id, Text, Transcription, Audio, Notes
	Word	Id, Text, Transcription
	Tag	Id, Description, Type

All the entities in the conceptual schema are mapped into the classes of the ontology, i.e. `actor`, for the derivation area, `region`, `province`, `town` for the geographical area, `document`, `tag`, `sentence`, `word` for the tagging area and `dialect` for the inter-area between the derivation and the geographical area. We map the `HasTagW` ternary relationship relating a tag to a word within a sentence in the ER schema (see Figure 2), into the `WordSentenceCollocation` class along with the properties relating it to the `Word`, `Sentence Collocation classes`. The same methodology has been followed to map the ternary relationship `IsSpeaker` into the `ActorSpeaksDocument` class. The hasSentence binary relationship of the ER schema along with its attribute position has been mapped into the class `Sentence Document Collocation`. All the attributes of the entities have been straightforwardly mapped into the classes of data types of the ontology by following the presented mapping rules.

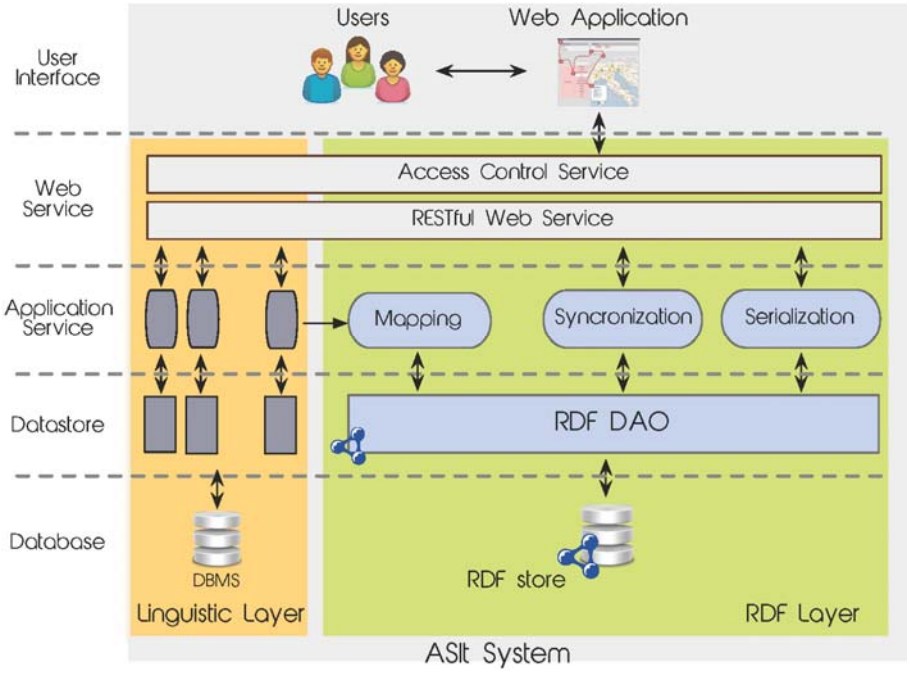
5 The architecture of the ASIt system

We present the architecture of the ASIt system composed of the *linguistic layer* and the *RDF layer*, as shown in Figure 3.

5.1 ASIt linguistic layer

The *linguistic layer* (Agosti et al., 2012) has been designed to be modular, thus reducing the dependency on a particular implementation of its constituting modules. It can be framed in four different levels: *database*, *datastore*, *application service*, and *web service*. The *database* level is constituted by a relational database, the schema of which is based on the ASIt conceptual model described in Section 4.2; the currently adopted DBMS is PostgreSQL.[25] The *datastore* is responsible for the persistence of the linguistic resources and provides an interface to store and access linguistic data. The *application service* is responsible for the interaction with the linguistic resources; it provides an API to perform operations on the resources – e.g. list sentences in a document, or list words in a sentence, and add tags to sentences and words. When linguistic resources are created or modified, the application service exploits the datastore API for the persistence of data. The *web service* provides functionalities to create, modify and delete resources, and gather their descriptions through appropriate HTTP requests based on a RESTful web service (Fielding and Taylor, 2002). This level is also responsible for access control which is necessary to preserve the quality of the data maintained in the ASIt database. Indeed, only allowed users can create or modify resources, whereas there is no restriction to access resource descriptions.

Figure 3 The architecture of the ASIt system in which we highlight the diverse constituting levels and the RDF layer (see online version for colours)



5.2 ASIt RDF layer

The *RDF layer* is responsible for persistence and access to RDF triples of linguistic data instantiated on the basis of the ontology. The RDF layer has been developed by exploiting the functionalities of the open source library Apache Jena.[26] Jena was adopted because of the variety of solutions for persistence of the RDF/S instantiation, the support of a number of RDF output formats, and the functionalities for reasoning with RDF and OWL data sources. A *mapping service* has been developed to instantiate the ontology starting from the data stored in the ASIt relational database. A request for creation, deletion or modification of a resource is processed by the linguistic layer that, through the proper module of the *application service*, allows the interaction with the resource and stores its new state. In parallel, by means of the *syncronisation service*, the RDF layer processes the request and updates the RDF triples instantiating the ASIt ontology. This service allows for the interaction with the *RDF datastore* which is responsible for the persistence of the RDF triples in the RDF store. Therefore, the operations required by resource creation, deletion or modification are performed in parallel for each request to guarantee the synchronisation between the relational database and the RDF store. When a request for accessing a resource is submitted to the system, the RDF *serialisation service* retrieves information on the requested resource from the RDF store and it returns the result in the requested output format (Di Buccio et al., 2015).

6 The ASIt geolinguistic linked open data set

The mapping between ASIt and the geolinguistic ontology allows us to expose the linguistic data of ASIt as a linked open data set whose details are reported in Table 4.

By exploiting the synchronisation services of the ASIt system, the ASIt geolinguistic linked open data set size grows proportionally to the size of the database. Table 5 reports the statistics about the evolution of the data in ASIt in the last two and a half years.

This data set has been exposed following the guidelines of Heath and Bizer (2011). As an example, we report how to access and browse a resource named 'Veneto' which is an instance of the class 'Region' by means of three URIs:

- http://purl.org/asit/resource/Region/Veneto
- http://purl.org/asit/data/Region/Veneto
- http://purl.org/asit/page/Region/Veneto

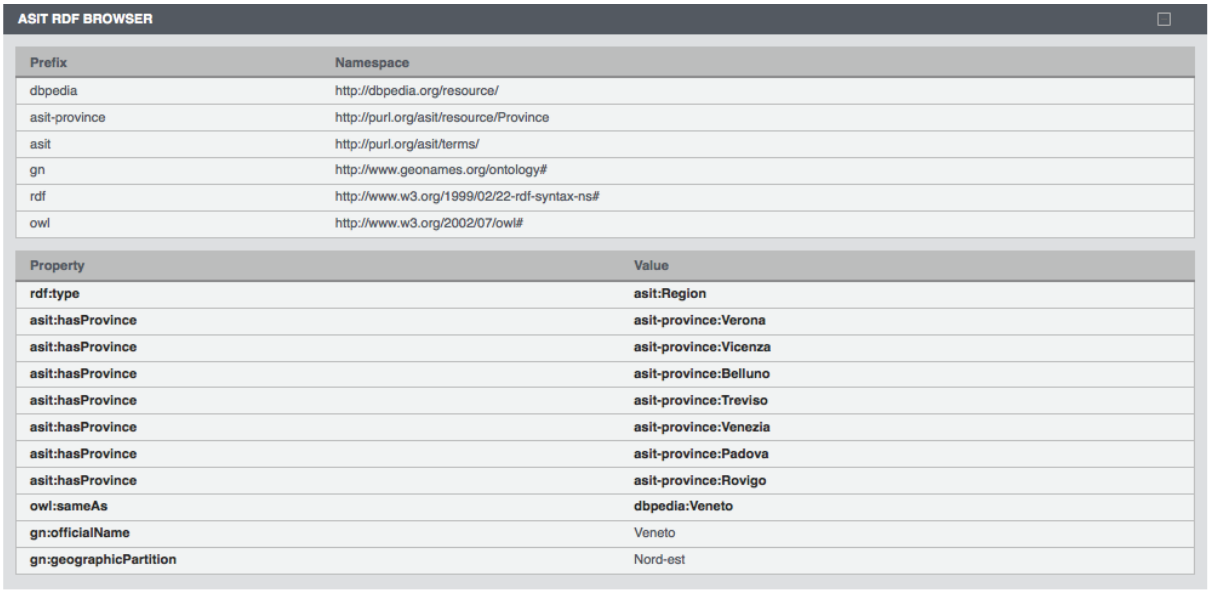
The first of the three URIs identifies the non-information resource[27] 'Veneto'. The second URI identifies the information counterpart of the same resource. If this URI is accessed by a web browser, the request is redirected to the third URI which refers to the HTML representation of 'Veneto'.

Currently, the ASIt data set is linked to DBpedia: the instances of the classes `region`, `province`, and `town` are linked to the corresponding instances of the dbpedia.org class `place` through the property `owl:sameAs`. Figure 4 shows an example of the RDF browser embedded in the ASIt system which is structured in two tables. The lower table reports the list of all the predicate-object pairs that has the resource 'Veneto' as subject. The upper table reports the list of the namespaces for the predicates and the subjects listed in the lower table. Figure 4 also reports the explicit link to the dbpedia resource http://dbpedia.org/page/Veneto. Hyperlinks to other resources (e.g. asit-province:Verona or dbpedia:Veneto) or ontology terms (gn:officialName) are in bold. When the user clicks on a link of a resource, the RDF browser tables are updated with predicates, objects and namespaces. In the event of an 'external' resource or/and ontology term, the user is redirected to the external service hosting its RDF description.

A SPARQL end point is provided at the URL:
http://purl.org/asit/rdf/sparql

and a GUI to submit queries to the ASIt Linguistic Linked Open Dataset is available at the URL:
http://purl.org/asit/rdf/sparqlGui

Figure 4 A screenshot of the ASIt RDF browser



7 A geolinguistic application

The system described in Section 5 allows data to be presented in different formats (currently RDF and JSON); besides access to the resources, functionalities to perform tag-based queries have been developed. Indeed, the objective of the ASIt project is to provide linguists with a system for investigating variations among closely related languages. We developed a graphical user interface on the top of the ASIt system that dynamically produces maps on the basis of the user request. The interface is available at the URL:

http://purl.org/asit/rdf/search

A screenshot is reported in Figure 5 where the circles highlight the steps performed by a user when performing a tag-based search. When the user accesses the search page, a table with the list of supported tag types is presented (see step 1). The table is dynamically populated by performing a query to the SPARQL end point, specifically:

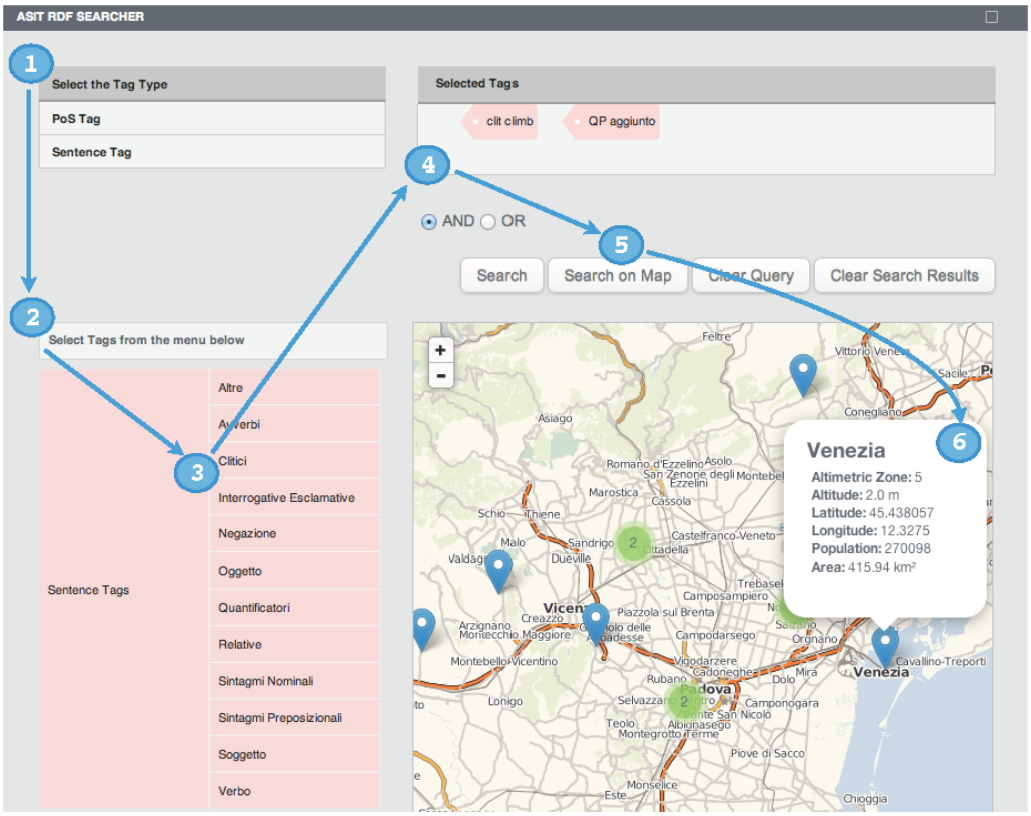
```
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX asit: <http://purl.org/asit/terms/>
PREFIX asit-tag: <http://purl.org/asit/resource/Tag/>

SELECT ?tagType ?label
{
        ?tagType rdfs:subClassOf asit:Tag.
        ?tagType rdfs:label ?label
}
```

When the user clicks on a tag type, a new table is displayed (see step 2) that reports a list of high-level tags for that tag type; in the example, the user selects the sentence tag type; this action results in the following SPARQL query:

```
PREFIX asit: <http://purl.org/asit/terms/>
SELECT DISTINCT
{
     ?parentTag ?label ?parentTag asit:isParentTag ?t.
     ?t asit:isTagSentence ?s.
     ?parentTag asit:tagDescription ?label
}
```

Tags of a given type are hierarchically organised in a tag tree. For instance, the root node of sentence tag three is the ‘Sentence Tags’ node; this node has 12 children that correspond to the 12 subsets in which the complete sentence tag set was divided by the linguists (the list is reported in the left column of the table labelled by 3). In order to display sentence tags at a finer level, the user can click on a specific high-level tag (in the example, ‘Clitici’) and a new table reporting the list of child tags is shown. Once the user clicks on a tag, the tag is added to the selected tag list (see step 4); the user can remove a tag from the list by clicking on it. Moreover, the user can select the Boolean constraint for the tag-based search: for instance, in Figure 5, the user was interested in all the sentences tagged with both ‘QP aggiunto’ and ‘clit climb’ sentence tags.

Figure 5 A screenshot of the ASIt RDF GeoSearch interface (see online version for colours)



Two different types of search are currently supported. The first search type returns the list of Italian sentences tagged by the requested sentence tags and that satisfy the specified Boolean constraint. Figure 6 depicts an example for this type of search; the corresponding SPARQL query is:

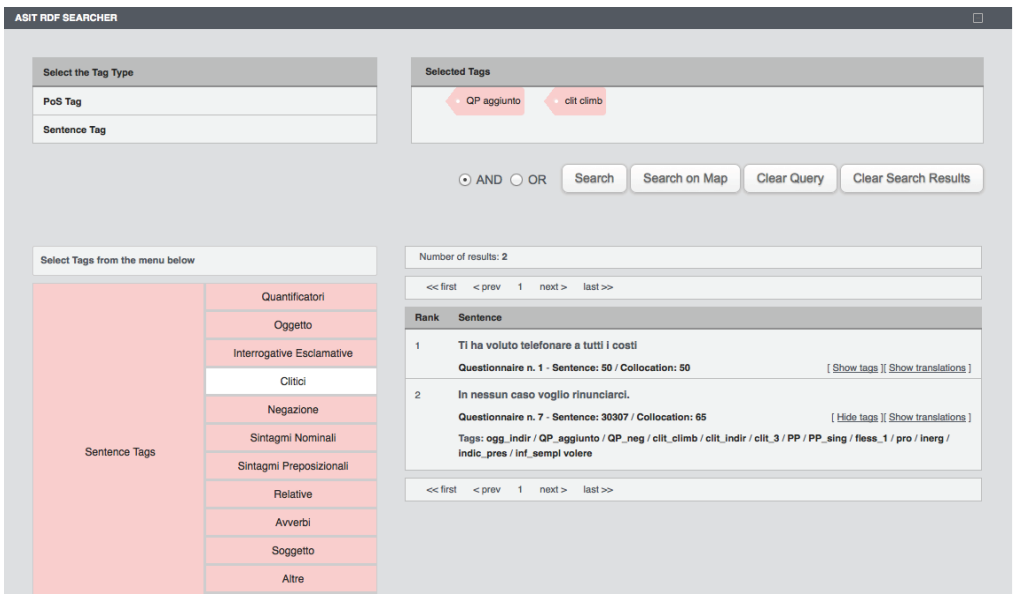
```
PREFIX asit: <http://purl.org/asit/terms/>
SELECT DISTINCT ?q ?s ?sp ?t
 WHERE {
    ?s asit:hasSentenceTag
      <http://purl.org/asit/resource/
            SentenceTag/QP_aggiunto>
    ?s asit:hasSentenceTag
       <http://purl.org/asit/resource/
            SentenceTag/clit_climb>
 ?q asit:hasSentence ?s
 ?s asit:sentence ?t
 ?qt asit:isTranslation ?q.
 ?sdc asit:sdcHasSentence ?s
 ?sdc asit:sdcHasDocument ?q
 ?sdc asit:sdcHasSentenceCollocation ?sc
   ?sc asit:sentencePosition ?sp
}  ORDER BY ?q ?sp
```

For each of the sentences in the result set, information on the questionnaire that contains that sentence, the sentence identifier and full text, and the sentence position within the questionnaire is reported; for instance, the second entry in the result set reported in Figure 6 refers to sentence 30,307 of the first questionnaire; the collocation of the sentence in the questionnaire is 65. The complete list of sentence tags for a sentence can be displayed by clicking on 'Show tags'. Questionnaire, sentence and tag information reported in the result list are links to the corresponding informative resources. In order to show the translation in the diverse Italian dialects for a sentence in the result list, the user can click on the link labelled with 'Show Translations'.

The second type of search is performed when the user clicks on the 'Search on Map' button. This type of search aims at satisfying the information need of a user searching for the geographic distribution of linguistic resources. For instance, the query submitted in Figure 6 retrieves all the Italian sentences tagged both by 'QP aggiunto' and 'clit climb'; for all these sentences, the system retrieves the locations for which a translation of the Italian sentence in the result set exists. The SPARQL query for the considered example is:

```
PREFIX asit: http://purl.org/asit/terms/
PREFIX geo: http://www.w3.org/2003/01/
geo/wgs84_pos#
PREFIX gn: <http://www.geonames.org/
ontology#>
SELECT DISTINCT ?t ?name ?lat ?long
?alt ?surface ?altimetricArea
?population
 WHERE
{
        ?s asit:hasSentenceTag
      <http://purl.org/asit/resource/
        SentenceTag/QP_aggiunto>.
 ?s asit:hasSentenceTag
     <http://purl.org/asit/resource/
         SentenceTag/clit_climb>.
?q asit:hasSentence ?s .
?qt asit:isTranslation ?q.
?qt asit:documentIsASD ?asd
 ?asd asit:asdHasDialect ?dia.
 ?dia asit:isSpokenIn ?t.
?t gn:officialName ?name.
?t geo:lat ?lat.
?t geo:long ?long.
?t geo:alt ?alt.
?t geo:population ?population.
?t asit:altimetricArea ?altimetricArea
?t asit:surface ?surface
}
ORDER BY ?lat ?long
```

Figure 6 A screenshot of the ASIt RDF search interface (see online version for colours)



The interface exploits the Leaflet javascript[28] library to visualise the results on a map; we adopted this library in order to achieve a complete open data approach: indeed the Leaflet library relies on *OpenStreetMap* data.[29] In order to display the locations in the result set, we used the Leaflet marker cluster plug-in. Each cluster is depicted as a circle; the number in the centre of the circle refers to the number of locations that belongs to the cluster. When the user clicks on a cluster, a zoom is performed in order to show all the locations within the cluster. Currently, location clustering is based on the default criterion implemented by the plug-in; future extension of our interface could support diverse clustering strategies, e.g. based on the features of the dialects spoken in the considered locations.

When the user clicks on the marker corresponding to a specific location, information on the location is displayed. For instance, step 6 in the considered example, the user clicked on the marker corresponding to the location 'Venezia' (i.e. Venice). A new and interactive visualisation of the ASIt linguistic linked data set has been proposed by Di Buccio et al. (2013).

8 Conclusions

Digital geolinguistic systems encourage collaboration between linguists, historians, archaeologists, ethnographers, as they explore the relationship between language and cultural adaptation and change. These systems can be used as instructional tools, presenting complex data and relationships in a way accessible to all educational levels. However, the heterogeneity of geolinguistic projects has been recognised as a key problem limiting the re-usability of linguistic tools and data collections.

In this paper, we propose an LOD approach to increasing the level of interoperability of geolinguistic applications and the re-use of the data. We defined an extensible ontology for geolinguistic resources based on the common ground defined by current European linguistic projects.

In this context, we studied the use case of a linguistic project named ASIt. We mapped the ASIt conceptual model into the geolinguistic ontology by taking into account several approaches that have been proposed in the literature. All the entities, relationships and attributes in the conceptual schema were mapped into the classes of the ontology by following the presented mapping rules. The architecture of the system which coordinates all the activities of the project was described in detail by dividing it in two main parts: (a) the linguistic layer, responsible for the activities carried out by linguists (creation, modification and search of linguistic resources); (b) the RDF layer, responsible for the persistence and access to RDF triples composing the publicly available linked open data set. By exploiting the newly defined synchronisation services, the ASIt geolinguistic linked open data set grows proportionally to the size of the database. Currently, the ASIt data set is linked to DBpedia showing how it is possible to extend it by establishing relationships with existing data sets.

One of the key points of this approach is the decoupling between the system which manages the data and the one which provides services over those data. In fact, we imagine the use of the geolinguistic linked open data set by third-party linguistic projects in order to enrich the data and build-up new services over them. As a concrete example, we presented a geolinguistic application build upon this data set which provides linguists with a system for investigating variations among closely related languages. Finally, we also developed a graphical user interface on top of this application that dynamically produces maps on the basis of the user requests.

References

Agosti, M., Alber, B., Di Nunzio, G.M., Dussin, M., Pescarini, D., Rabanus, S. and Tomaselli. A. (2011) 'A digital library of grammatical resources for European dialects', in Agosti, M., Esposito, F., Meghini, C. and Orio, N. (Eds): *Communications in Computer and Information Science*, Springer, pp.61–74.

Agosti, M., Alber, B., Di Nunzio, G.M., Dussin, M., Rabanus, S. and Tomaselli, A. (2012) 'A curated database for linguistic research: The test case of Cimbrian varieties', in Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J. and Piperidis, S. (Eds): *European Language Resources Association (ELRA)*, pp.2230–2236.

Agosti, M., Benincà, P., Di Nunzio, G.M., Miotto, R. and Pescarini, D. (2010) 'A digital library effort to support the building of grammatical resources for Italian dialects', in Agosti, M., Esposito, F. and Thanos, C. (Eds): *Communications in Computer and Information Science*, Springer, pp.89–100.

Akmajian, A., Demers, R.A., Farmer, A.K. and Harnish, R.M. (2010) *Linguistics: An Introduction to Language and Communication*, 6th ed., the MIT Press, Cambridge, MA.

Alvar Lòpez, M. and Nuño Alvarez, M.P. (1981) 'Un ejemplo de atlas lingüìstico automatizado', *LEA: Lingüìstica española actual*, Vol. 3, No. 2, pp.359–374.

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S. and Aumueller, D. (2009) 'Triplify: light-weight linked data publication from relational databases', *Proceedings of the 18th International Conference on World Wide Web*, ACM Press, pp.621–630.

Bael, J.C., Corrigan, K.P. and Moisl, H.L. (2006) *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*, Palgrave Macmillan.

Benincà, P. and Poletto, C. (2007) 'The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects', *Nordlyd*, Vol. 34, pp.35–52.

Bird, S., Klein, E. and Loper, E. (2009) *Natural Language Processing with Python*, O'Reilly Media.

Breton, R.J.L. and Schiffman, H.F. (1991) *Geolinguistics: Language Dynamics and Ethnolinguistic Geography*, University of Ottawa Press, Canada.

Buneman, P., Cheney, J., Chiew Tan, W. and Vansummeren, S. (2008) 'Curated databases', Lenzerini, M. and Lembo, D. (Eds): *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'2008)*, ACM Press, pp.1–12.

Chiarcos, C. (2012) 'Interoperability of corpora and annotations', in Chiarcos, C., Nordhoff, S. and Hellmann, S. (Eds): *Linked Data in Linguistics*, Springer Berlin, Heidelberg, pp.161–179.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J. and Stede, M. (2008) 'A flexible framework for integrating annotations from different tools and tag sets', *TAL*, Vol. 49, No. 2, pp.217–246.

Di Buccio, E., Di Nunzio, G.M. and Silvello, G. (2012) 'A system for exposing linguistic linked open data', *Research and Advanced Technology for Digital Libraries – International Conference on Theory and Practice of Digital Libraries (TPDL'2012)*, Papho, Cyprus, pp.172–178.

Di Buccio, E., Di Nunzio, G.M. and Silvello, G. (2013) 'A curated and evolving linguistic linked dataset', *Semantic Web*, Vol. 4, No. 3, pp.265–270.

Di Buccio, E., Di Nunzio, G.M. and Silvello, G. (2013) 'A geolinguistic web application based on linked open data', *SIGIR'13: 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.1101–110.

Farrar, S. and Langendoen, T. (2003) 'A linguistic ontology for the semantic web', *Glot International*, Vol. 7, No. 3, pp.97–100.

Fielding, R.T. and Taylor, R.N. (2002) 'Principled design of the modern web architecture', *ACM TOIT*, Vol. 2, pp.115–150.

Haspelmath, M., Dryer, M.S., Gil, D. and Comrie, B. (2005) *The World Atlas of Language Structures*, Oxford University Press, UK.

Heath, T. and Bizer, C. (2011) *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool Publishers.

Hoch, S. and Hayes, J.J. (2010) 'Geolinguistics: the incorporation of geographic information systems and science', *The Geographical Bulletin*, Vol. 51, No. 1, pp.23–36.

Kilgarriff, A. (2007) 'Googleology is bad science', *Computational Linguistics*, Vol. 33, No. 1, pp.147–151.

Klyne, G. and Carroll, J.J. (2004) *Resource Description Framework (RDF): Concepts and Abstract Syntax*, Technical Report, W3C.

Kunst, J.P. and Wesseling, F. (2011) 'The Edisyn search engine', *Language Variation Infrastructure*, Vol. 3, No. 2, pp.63–74.

Lameli, A., Kehrein, R. and Rabanus, S. (2010) *Language and Space: Language Mapping: An International Handbook of Linguistic Variation*, Walter de Gruyter.

Myroshnichenko, I. and Murphy, M.C. (2009) 'Mapping ER schemas to OWL ontologies', *Proceedings of the IEEE International Conference on Semantic Computing*, pp.324–329.

Odijk, J. (2010) 'The CLARIN-NL project', *LREC*, European Language Resources Association.

Spärck Jones, K. (2007) 'Computational linguistics: what about the linguistics?', *Computational Linguistics*, Vol. 33, No. 3, pp.437–441.

Xie, Y., Aristar-Dry, H., Aristar, A., Lockwood, H., Thompson, J., Parker, D. and Cool, B. (2009) 'Language and location: map annotation project – a GIS-based infrastructure for linguistics information management', *International Multiconference on Computer Science and Information Technology (IMCSIT'09)*, pp.305–311.

Notes

1 http://en.wikipedia.org/wiki/Georg_Wenker

2 http://www.wals.info/

3 http://www.dialectsyntax.org/

4 http://tools.ietf.org/html/rfc3986

5 http://www.w3.org/RDF/

6 A curated database is a database the content of which has been collected by a great deal of human effort.

7 http://www.language-archives.org

8 http://linguistics-ontology.org/

9 http://www.clul.ul.pt/

10 http://www.tekstlab.uio.no/nota/scandiasyn/

11 http://www2.hu-berlin.de/vivaldi/tagung/index.html

12 http://www2.hu-berlin.de/vivaldi/

13 http://ald1.sbg.ac.at/

15 http://stel.ub.edu/labfon/amper/

15 http://www.atlas-alltagssprache.de

16 http://www2.unine.ch/dialectologie/page-8174.html

17 http://www.soundcomparisons.com

18 http://www.geolectos.com/

19 http://llmap.org/

20 http://www.w3.org/DesignIssues/LinkedData.html

21 http://www.w3.org/TR/owl-features/

22 RDF assumes that any instance of a class may have an arbitrary number (zero or more) of values for a particular property. As an extension of RDF/S, OWL allows us to specify the maximum number of occurrences of a class within a property.

23 http://www.w3.org/TR/rdf-concepts/

24 http://www.w3.org/TR/xmlschema-2/

25 http://www.postgresql.org/

26 http://incubator.apache.org/jena

27 "All the resources we find on the traditional document Web, such as documents, images, and other media files, are information resources. All 'real-world objects' that exist outside of the web are non-information resources" (Heath and Bizer, 2011)

28 http://leafletjs.com/

29 http://www.openstreetdata.org/